

Поисковая система
Avalanche Personal



Руководство пользователя

«Интерруссофт», 2000-2004

Содержание

Содержание	2
Введение	4
Общая информация	5
Системные требования	5
Установка и удаление	6
Запуск программ Spider и Avalanche	6
Работа с Avalanche	7
Интерфейс программы	7
Главное меню	8
Панель инструментов	9
Работа со списком собранных документов	10
Просмотр и редактирование собранных документов	11
Просмотр собранных документов	12
По рубрикам.....	13
По источникам	13
По ключевым словам.....	14
Работа с каталогом «Избранное»	15
Добавление документов	16
Экспорт документов	17
Общая информация о системе экспорта	17
Выполнение экспорта.....	18
Удаление и добавление настроек экспорта	20
Настройка экспорта в формате HTML.....	22
Настройка экспорта в формате XML	25
Просмотр статистики поступления	26
Настройка базы данных	26
Настройка прокси-сервера	27
Работа с поисковым агентом (Spider)	29
Функционирование программы	29
Главное меню	30
Рабочий цикл программы	32
Настройка системы	32
Формирование списка сайтов.....	32
Работа со списком сайтов	33
Настройка ссылок	35
Настройка рубрик	43
Определение политики сбора	46
Настройка базы данных	46
Сохранение настроек.....	47
Сбор сообщений	48
Контроль за ходом выполнения программы	49

Просмотр результатов сбора	50
Горячие клавиши	51
<i>Словарь терминов.....</i>	52

Введение

Система предназначена для сбора сообщений, циркулирующих в сети Интернет. Сообщения извлекаются из сети и сохраняются в локальной базе данных на диске. Пользователю предоставляется удобный инструмент доступа к данным, который позволяет производить поиск сообщений по различным атрибутам, по ключевым словам, выделять и структурировать нужные сообщения и производить экспорт в переносимом формате, удобном для публикации в Сети.

Система обеспечивает:

- Оптимизированный интеллектуальный сбор информации в сети;
- Оперативное информирование о ходе и результатах выполнения работы;
- Эффективный анализ собранных данных на предмет выявления и вычленения целостных сообщений;
- Выявление атрибутов сообщений;
- Получение информации из поисковых систем Интернета.
- Хранение найденной информации в базе данных;
- Фильтрация и рубрикация полученных сообщений в соответствии с заданным перечнем тем;
- Просмотр, ручное добавление, удаление и редактирование сообщений;
- Поиск сообщений по рубрикам, источникам, ключевым словам и времени публикации.
- Экспорт сообщений в формате HTML, удобном для просмотра и готовом для публикации в сети Интернет виде.
- Организация и структурирование нужных Вам сообщений в "Избранном"
- Автоматический подсчет статистики поступающих сообщений и представление ее в виде графика.

Общая информация

Здесь приведены основные требования к программной среде и оборудованию, выполнение которых необходимо для эффективного применения системы (**Системные требования**), а также указан порядок инсталляции, запуска и удаления программы (**Инсталляция и удаление программы, Запуск программы**).

Системные требования

Продукт поддерживает операционные системы Windows 9x/Me/2000/XP/Server 2003.

Рабочая станция, на которой устанавливается продукт, должна соответствовать следующим требованиям:

- Процессор – класса не ниже, чем Pentium III 500. Рекомендуется процессор класса Pentium III 1000 Mhz.
- Оперативная память – 128 Мб. Рекомендуется – 256 Мб.
- Видеосистема – разрешение не менее, чем 800x600 с глубиной цвета не менее, чем 16 бит цветов.
- Место на жестком диске – 70 Мб + дополнительный объем для хранения поступающих сообщений
- Мышь, или аналогичное устройство ввода.

Программные требования:

- Операционная система: Windows 9x/Me/2000/XP/Server 2003 (рекомендуется Windows 2000/XP/Server 2003)
- Наличие подключения к Интернет или интранет.
- Поддержка протоколов TCP/IP, HTTP.
- Наличие веб-браузера Internet Explorer, или другого веб-браузера. Требуется для просмотра экспортируемых сообщений.

Установка и удаление

Установка

Установка осуществляется запуском файла setup.exe, находящимся на инсталляционном диске. Далее следуйте инструкциям программы установки.

Удаление программы

Удаление продукта осуществляется стандартными средствами Windows:

Start->Settings->Control Panel->Add or Remove Programs,

или

Пуск->Настройки->Панель управления->Установка и удаление программ.

Примечание: После удаления в папке где находилась программа может остаться папка с базой данных (Dataset). Если Вы захотите в будущем продолжить использование продукта с накопленными данными и настройками, сохраните эту папку.

Запуск программ Spider и Avalanche.

Запуск продукта осуществляется стандартными средствами Windows:

Программа - поисковик (осуществляет просмотр собранных сообщений):

кнопка **Старт(Start)->Программы(Programs)->Avalanche->Avalanche**

Программа - поисковый агент (осуществляет сбор новостей в Сети):

кнопка **Старт(Start)->Программы(Programs)->Avalanche->Spider**

Работа с Avalanche

Программа-поисковик Avalanche предназначена для поиска, просмотра и организации сообщений (новостей), извлеченных из Сети поисковым агентом Spider. Она является основным инструментом вашей работы с системой.

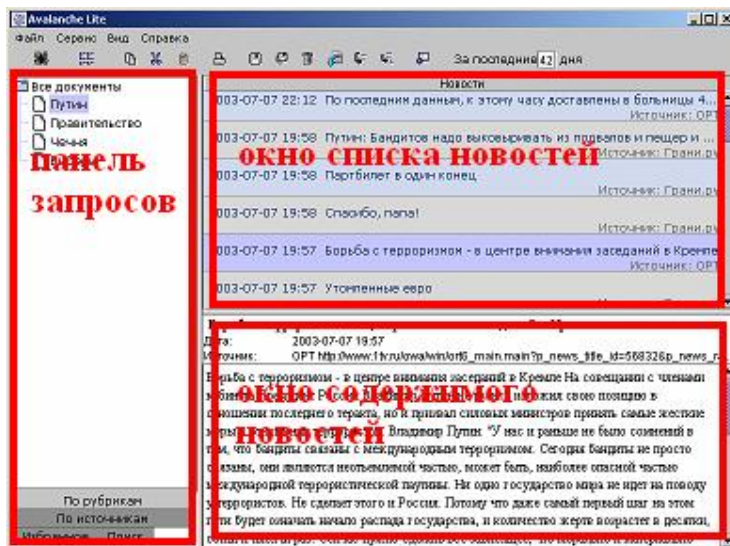
Доступ к программе-поисковику осуществляется через меню *Вид->Просмотр результатов* поискового агента, или запуском программы Avalanche через меню Пуск (Start)->Программы (Programs)->Avalanche->Avalanche.

После общего ознакомления с интерфейсом программы, получить минимум необходимой информации для работы с системой можно ознакомившись с документом "Быстрый старт".

Интерфейс программы

Окно программы просмотра содержит:

- панель запросов, в которой пользователь выбирает критерий выборки новостей для просмотра
- окно списка новостей
- окно просмотра новостей
- главное меню
- панель инструментов



Для того чтобы приступить к работе прямо сейчас, воспользуйтесь документом "Быстрый старт".

Главное меню

Меню "Файл":

- | | |
|------------------------------|--|
| Поменять рабочую базу | - создание базы данных программы или изменение её расположения |
| Выход | - завершение работы программы |

Меню "Сервис":

- | | |
|---------------------------------------|---|
| Открыть окно поискового агента | - вызывает окно поискового агента "Паук" для настройки и сбора сообщений в Сети |
| Добавить документ вручную | - вызывает окно ручного добавления документов |
| Экспорт | - вызывает окно экспортирования документов |
| Настройка прокси | - вызывает диалог настройки программы-паука для работы с прокси сервером |

Меню "Вид":

- | | |
|--------------------------------|---|
| Темы | - позволяет пользователю выбрать стиль внешнего вида приложения на свой вкус |
| Настройка рубрик | - вызывает окно настройки рубрик (добавление, удаление, изменение булева фильтра) см. Настройка рубрик поискового агента. |
| Выделять ключевые слова | - включает/выключает режим выделения красным цветом ключевых слов рубрики в тексте сообщения. |

Меню "Справка":

- | | |
|---------------------------|---|
| Справка по системе | - вызывает эту систему справки |
| О программе | - отображает информацию о программе и ее создателях |

Панель инструментов

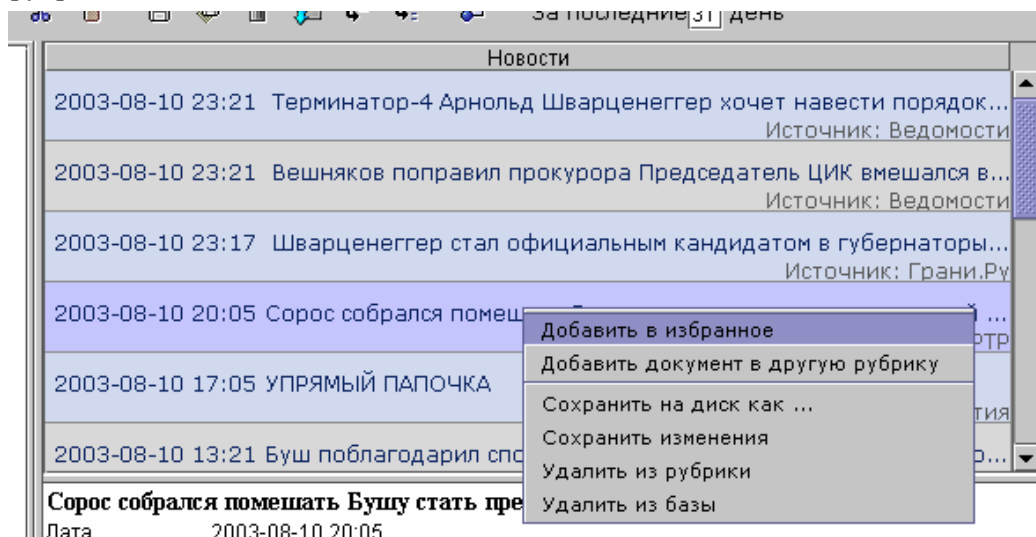


Панель инструментов обеспечивает быстрый доступ к некоторым функциям программы:

- вызывает окно поискового агента "Паук" для настройки и сбора сообщений в Сети.
- вызывает окно ручного добавления документов
- вызывает окно настройки рубрик. см. Настройка рубрик поискового агента.
- копировать выделенный фрагмент текста в буфер обмена
- вырезать выделенный фрагмент текста в буфер обмена
- вставить фрагмент текста, находящийся в буфере обмена в текущую позицию курсора
- сохранить текущий документ на диск
- сохранить изменения, сделанные в текущем документе пользователем в базу данных
- удалить текущий документ из базы данных
- удалить текущий документ из рубрики (только при просмотре по рубрикам)
- добавить текущий документ в каталог "Избранное"
- добавить текущий документ в другую рубрику (только при просмотре по рубрикам)
- просмотр статистики поступления документов по данной рубрике.
- указать возраст документов (в сутках), с которыми будет работать пользователь.

Работа со списком собранных документов

Список документов в правой верхней части окна представляет собой результаты выборки документов по определенному критерию, например по указанной рубрике:



Для каждого документа указывается его дата публикации на новостной ленте, название и источник сообщения.

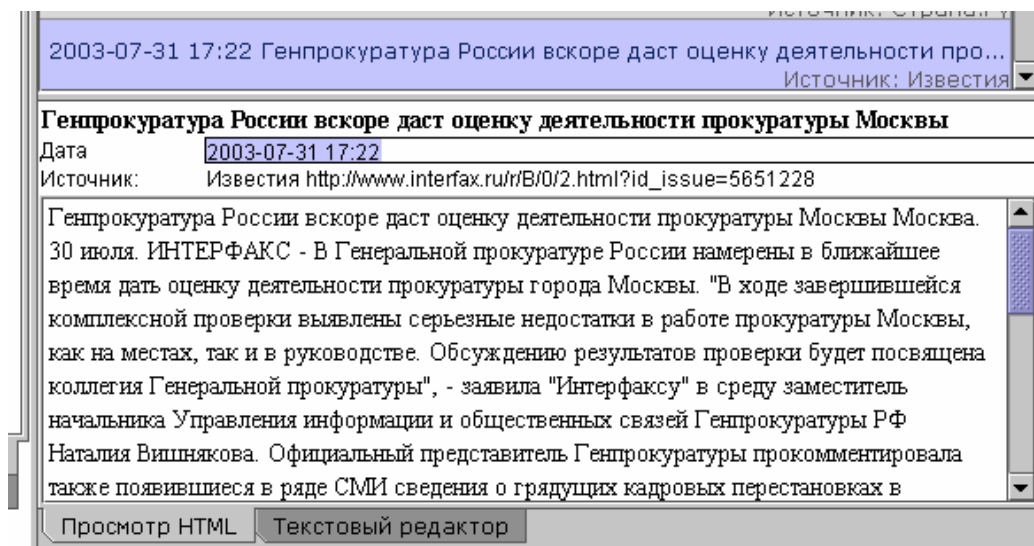
Для выбора конкретного сообщения нужно щелкнуть по нему левой кнопкой мыши. Для выбора нескольких сообщений одновременно, держите нажатой клавишу Ctrl и левой кнопкой мыши выделяйте сообщения. Щелчок правой кнопкой мыши вызывает контекстное меню для данного сообщения, которое позволяет:

- добавить документ в пользовательский каталог "Избранное"
- добавить документ в другую рубрику, если вы считаете что документ туда не попал по ошибке. При этом появляется окно со списком рубрик. Выберите нужную Вам и нажмите ОК. (Этот пункт доступен только в режиме просмотра по рубрикам.)
- "сохранить на диск как..." позволяет сохранить выделенное сообщение на ваш диск в указанную директорию в формате HTML, при этом рекомендуется к имени файла добавлять расширение ".html" или ".htm"
- сохранить изменения в текущем документе, если он был редактирован пользователем. См. редактирование документов.
- удалить из текущей рубрики выделенный документ, если вы считаете, что он туда попал ошибочно. (Этот пункт доступен только в режиме просмотра по рубрикам.)

- удалить из базы выделенный документ, если он Вам больше не понадобится. После удаления документ восстановлению не подлежит.

Просмотр и редактирование собранных документов

После выбора документа из списка, его содержимое отображается в окне справа внизу:



В окне отображается:

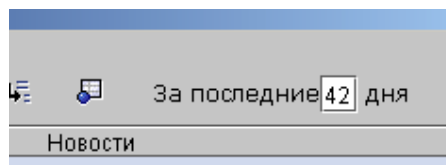
- Название сообщения (заголовок новости). Его можно редактировать, сделав по нему двойной щелчок левой кнопкой мыши. При этом отображается курсор редактирования. Произведите необходимые изменения, затем нажмите Enter.
- Дата поступления сообщения на ленту (извлекается со страницы источника). Редактируется аналогично названию сообщения.
- Источник сообщения (гиперссылка в Сети). Для того чтобы перейти на страницу источника сообщения в Сети, достаточно нажать левой кнопкой мыши по источнику сообщения в заголовке. При этом откроется страница с источником сообщения в интернет-браузере.
- Текст сообщения (вкладка "Просмотр HTML") отображается отформатированным, в конце сообщения отображаются картинки, ассоциированные с сообщением. Вы увидите картинки, только если ваш компьютер в текущий момент подключен к сети Интернет.
- Текстовый редактор сообщения позволяет редактировать исходный текст сообщения, хранящийся в базе данных. Текст хранится в формате HTML, так что редактор отображает все конструкции этого языка (элементы форматирования, гиперссылки...). В большинстве случаев для

редактирования сообщения не требуется знание HTML. После редактирования необходимо сохранить текущие изменения, нажав кнопку "сохранить изменения" на панели инструментов или в контекстном меню списка документов.

Просмотр собранных документов

На лентах новостей ежедневно появляется большое количество сообщений, таким образом, чтобы найти и выделить нужные Вам документы необходимо выработать критерий отбора. Основными критериями являются булевы фильтры, заданные в рубриках. Также в качестве критерия может выступать источник сообщения (определенная новостная лента). Безусловно, на все это нужно накладывать условия на возраст сообщений, чтобы не просматривать наряду с сегодняшними новости годичной давности.

1. Критерий возраста сообщений:



указывается на панели инструментов и начинает работать при выборе любого пункта (рубрики или источника) из списка слева.

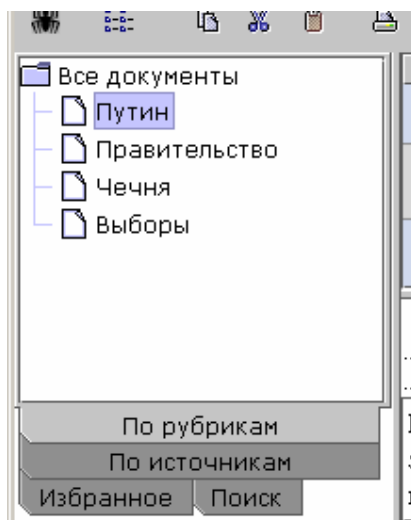
2. Выбор критерия поиска новостей следует указать, выбрав вкладку слева внизу под списком рубрик или источников:

- вкладка "Рубрики" показывает список рубрик для просмотра документов по рубрикам. При этом один документ может принадлежать нескольким рубрикам. Также контекстное меню в списке рубрик позволяет просмотреть статистику сообщений для выбранной рубрики.
- вкладка "Источники" отображает список источников, при щелчке на которых появляется список документов, с них поступивших.
- вкладка "Поиск" служит для поиска сообщений из выбранных источников и диапазона дат по заданным ключевым словам.

Вкладка "Избранное" не является критерием отбора сообщений, а служит своеобразным хранилищем выбранных пользователями документов. Если вас заинтересовал какой-либо документ, вы можете поместить его в "Избранное" (с помощью контекстного меню списка документов), чтобы в дальнейшем быстро найти его. В избранном допускается создание пользователем иерархической структуры для классификации помещенных туда документов.

По рубрикам

Просмотр документов по рубрикам отображает документы, как они были автоматически классифицированы согласно булевому фильтру рубрики во время сбора сообщений. Так как булев фильтр не обеспечивает анализа документов по смыслу (т.е. семантического анализа), то иногда могут попадаться сообщения, которые удовлетворяют фильтру, но все же относятся к другой рубрике. В этом случае можно перенести документ в другую рубрику, затем удалить его из текущей рубрики. Эти операции можно выполнить, выделив документ в списке документов, с помощью контекстного меню, или панели инструментов.



На вкладке отображается дерево рубрик. При нажатии на корень дерева левой кнопкой мыши в списке документов отображаются все собранные документы за последнее указанное количество дней, даже те, которые не принадлежат не одной из рубрик. При выборе определенной рубрики отображаются только те документы, которые ей принадлежат. Один документ может принадлежать нескольким рубрикам одновременно.

При нажатии правой кнопки мыши на одном из элементов дерева отображается контекстное меню:

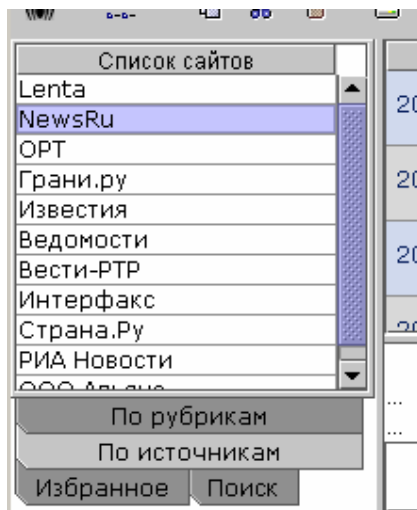
- контекстное меню корневого элемента дерева "Все документы" позволяет обновить список рубрик. Этот пункт полезен, если вы изменили список рубрик из программы-поискового агента Spider.
- контекстное меню любой рубрики позволяет получить статистику поступления сообщений в данную рубрику за последние 7 дней.

По источникам

Каждый документ был получен поисковым агентом из определенного источника в Интернете, или введен пользователем вручную. Просмотреть все

документы, полученные за последнее указанное количество дней из определенного источника позволяет вкладка "По источникам".

- Выберите вкладку "По источникам"
- Выберите один из источников из списка
- В списке документов отобразятся все документы, полученные из этого источника.



После всех источников можно найти кнопку "Обновить список", нажатие на которую обновляет список источников. Эта функция полезна, если вы только что изменили список источников в программе-поисковом агенте Spider.

По ключевым словам

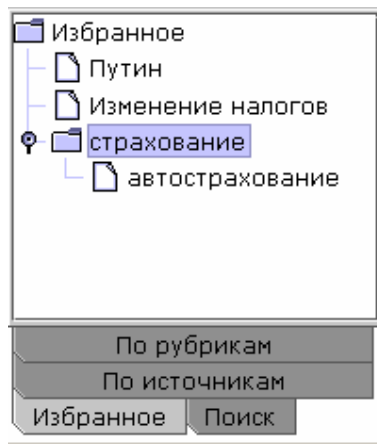
Окно поиска позволяет производить выборку новостей по ограничениям, заданным на текст, источник и дату новости:

1. Введите список ключевых слов через пробел. Например "банк курс евро". Регистр букв учитывается;
2. выберите список источников из всплывающего окна, отметив нужные галочкой. По умолчанию выбраны все источники;
3. введите диапазон дат в формате ГГГГ-ММ-ДД. Например "2003-04-25" - 25 апреля 2003 года. По умолчанию выбираются сообщения любого возраста. Возраст сообщений указанный на панели инструментов не учитывается.
4. нажмите "Поиск".

Соответствующие критериям новости появятся в списке справа. Работа с ними аналогична работе с обычной выборкой по рубрике или источнику.

Работа с каталогом «Избранное»

Каталог "Избранное" служит для отбора интересующих вас сообщений и их классификации по вашему вкусу.



Доступ к "Избранному" осуществляется через соответствующую вкладку на левой панели.

Документы хранятся в "избранном" подобно файлам в файловой системе. Существоют папки, в которых могут храниться документы и другие папки.

Для добавления новой папки в "избранное":

1. Щелкните правой кнопкой мыши по любой папке в "избранном", в которой вы хотите создать папку, либо по корневому элементу "Избранное".
2. В контекстном меню выберите пункт "добавить папку".
3. Введите название новой папки и нажмите ОК.

Для удаления папки в "избранном":

1. Выделите папку, которую вы хотите удалить и щелкните по ней правой кнопкой мыши.
2. В контекстном меню выберите пункт "удалить папку". При удалении папки удаляются все находящиеся в ней документы. Для удаления папок в которых находятся другие папки, сначала удалите их, а затем родительскую папку. Удаление папок со всеми подпапками запрещено.

Для добавления документа в папку "Избранное":

1. Выберите документ в списке документов.
2. Нажмите на него правой кнопкой мыши и в контекстном меню выберите "Добавить в избранное", или нажмите кнопку "добавить в избранное" на панели инструментов.
3. В появившемся окне выберите папку в "избранном", в которую вы хотите добавить документ и нажмите ОК.

Для удаления документа из "избранного":

1. Выберите одну из папок в "избранном" в окне слева.
2. выделите документ, которых хотите удалить справа.
3. Нажмите на нем правой кнопкой мыши и в контекстном меню выберите "Удалить файл". Документ восстановлению не подлежит.

Добавление документов

Иногда требуется ввести документ в базу данных вручную, если его источник находится не на новостной ленте в Интернете, или не подлежит настройке для автоматического сбора.

Для ручного добавления документа выберите пункт меню Сервис->добавить

документ вручную, или соответствующую кнопку на панели инструментов. После этого появится следующий диалог:

- Введите заголовок документа
- Введите дату получения документа (не обязательно текущую) в формате ГГГГ-ММ-ДД.

- Введите время получения документа (не обязательно текущее) в формате ЧЧ:ММ
- Название источника
- URL источника в интернете (если есть)
- Нажмите "редактировать", чтобы определить принадлежность документа рубрикам. (нужно снять крестики в рубриках, которым он должен принадлежать).
- Введите текст документа (например с помощью операций с буфером обмена - Ctrl-C - копировать текст, Ctrl-V - вставить текст).
- Нажмите "Добавить".

Экспорт документов

Общая информация о системе экспорта

Система поиска и сбора сообщений в сети Интернет Avalanche позволяет экспортировать сообщения, находящиеся в ее базе данных в файловую систему операционной среды. Экспорт производится в файлы текстового формата согласно настраиваемому шаблону экспорта.

Из базы данных экспортируются сообщения указанного возраста, то есть те, которые доступны в Avalanche при активной вкладке "По рубрикам". Возраст сообщений указывается в сутках, соответствующий элемент управления находится на панели инструментов («За последние <N> дней»).

После экспорта, в директории, куда был произведен экспорт находятся следующие группы файлов:

А) один файл *index.html* – индексная страница всех экспортируемых сообщений

Б) один файл *!index.html* – альтернативная индексная страница

В) некоторое количество файлов (по количеству экспортируемых сообщений) с именами **<числовой идентификатор сообщения>.html** (без угловых скобок).

Экспорт каждой группы файлов производится по заданным шаблонам, по одному для каждой группы. Соответственно, имеются три шаблона, которые указываются в настройках экспорта (меню Сервис/Настройка экспорта) на вкладках *Первая страница* (файл *index.html*), *Дополнительная страница* (файл *!index.html*), и *Новостные страницы* (группа файлов В),.

Перед экспортом сообщений необходимо задать значения переменных, имена которых присутствуют в шаблонах и задаются пользователем при настройке шаблонов на вкладке *Переменные пользователя* в настройках экспорта.

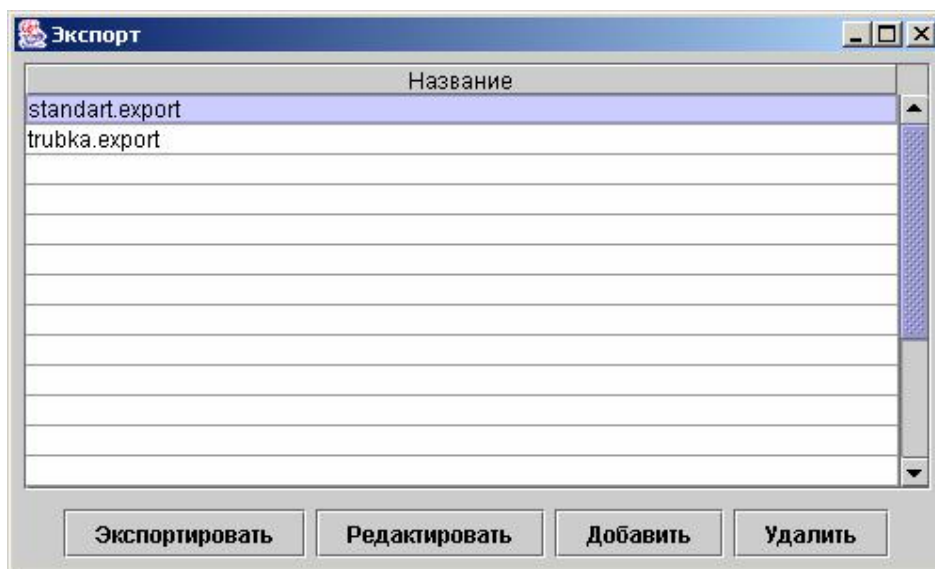
Все настройки экспорта хранятся в директории, где находится база данных в файлах с расширением *.export*. Файлы могут быть скопированы, или сохранены в другое место без потери информации. Для того, чтобы файл с настройками появился в списке, необходимо и достаточно, чтобы он находился в папке базы данных, которая в данный момент активна и имел расширение «*.export*».

Шаблоны страниц экспорта позволяют использовать информацию о рубриках (название и группировка сообщений по рубрикам) и сообщениях (заголовок, подзаголовок, дата, время, источник, ссылка на источник, текст сообщения).

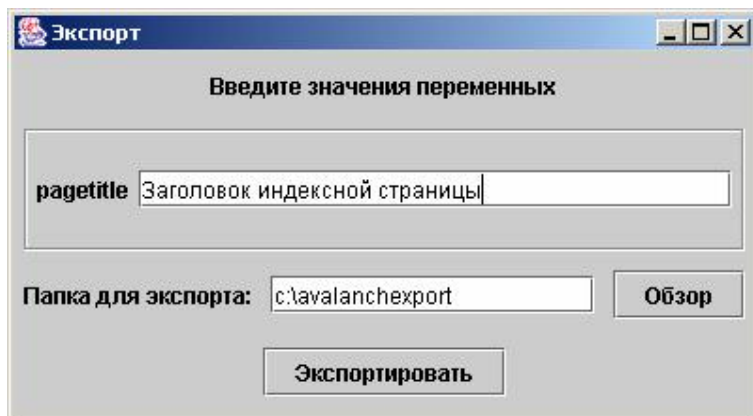
Выполнение экспорта

Для того, чтобы *произвести экспорт сообщений* нужно выполнить следующие действия:

1. Запустить программу-поисковик *Avalanche*.
2. Указать на панели инструментов количество дней, то есть возраст экспортируемых сообщений.
3. Вызвать меню *сервис/экспорт*:



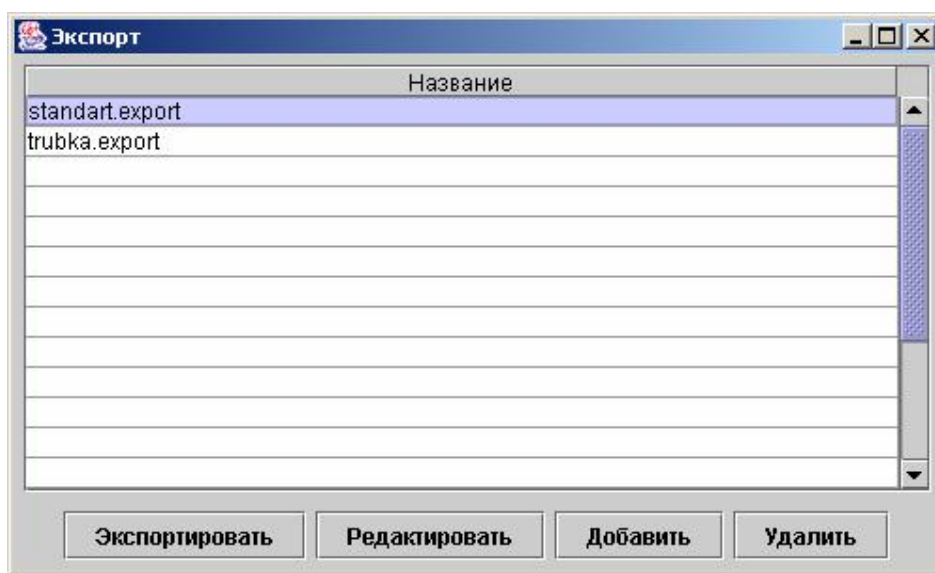
4. Выбрать необходимую Вам настройку экспорта из доступных и нажать «Экспортировать».
5. В диалоге экспорта указать значения пользовательских переменных и папку, в которую будет произведен экспорт:



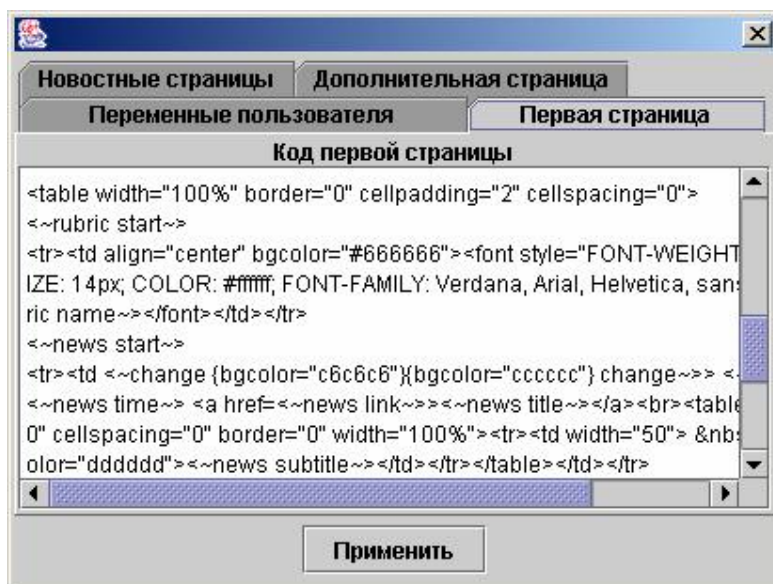
6. Нажать кнопку *Экспортировать*.
7. После этого в указанной папке появятся файлы index.html, !index.html и .html-файлы с сообщениями.

Для того, чтобы *настроить экспорт сообщений*:

1. Запустить программу-поисковик Avalanche.
2. Вызвать пункт меню *сервис/экспорт*.
3. Выбрать файл из списка настроек, который вы хотите изменить и нажать «Редактировать»:



4. Подготовить и вставить текстовые шаблоны хотя бы одного индексного файла («Первая страница» и «Дополнительная страница») и шаблон файла с сообщением («Новостные страницы»):



5. Указать переменные пользователя, используемые в шаблонах и, возможно, их описание.
6. Нажать кнопку «Применить». Настройки сохранятся в соответствующем файле, который Вы выбрали в п.3.

Удаление и добавление настроек экспорта

Для того чтобы *создать новую настройку* экспорта, необходимо:

1. Запустить программу-поисковик Avalanche.
2. Вызвать пункт меню сервис/экспорт.
3. Нажать кнопку «добавить»
4. Ввести название настройки латинскими буквами и нажать «Ок»
5. В список добавится только что созданный Вами файл с указанным Вами именем и расширением «.export». Файл с таким же именем появится в каталоге, где находится активная база.

6. Созданный файл настройки первоначально не содержит никаких шаблонов. Вам нужно сформировать их по потребностям.

Для того, чтобы **скопировать настройку**, или **использовать существующий файл с настройкой** экспорта:

1. Откройте папку, где расположена используемая база данных (можно посмотреть в меню «Файл/Поменять рабочую базу – путь к базе данных»).
2. Если необходимо скопировать существующую настройку – скопируйте файл с настройкой средствами операционной системой и присвойте ему нужное имя, сохранив расширение «.export».
3. Если необходимо использовать существующий файл с настройкой (например, если вы получили файл по электронной почте) – также скопируйте его в директорию БД и переименуйте средствами ОС, используя расширение «.export»
4. При следующем открытии диалога сервис/экспорт Вы увидите добавленную настройку.

Для **удаления настройки**:

1. Запустить программу-поисковик Avalanche.
2. Вызвать пункт меню сервис/экспорт. Нажать кнопку «удалить». При этом удаляется файл настройки без возможности восстановления.

Настройка экспорта в формате HTML

Для экспорта в формате HTML необходимо задать три шаблона, которые содержат инструкции для препроцессора системы экспорта и промежуточный HTML-код.

При экспорте, на вход препроцессору подаются шаблоны и данные из базы данных сообщений, на выходе получается описанный выше набор .HTML-файлов. Команды препроцессора имеют следующий вид:

<~команда~>

В качестве *команды* может выступать предопределенная конструкция, описание которых приведено ниже, либо произвольное английское сочетание букв, которое является пользовательской переменной и должно быть зарегистрировано в списке пользовательских переменных в настройках экспорта. Значение переменной вводится пользователем непосредственно при экспорте.

Предопределенные конструкции в шаблонах индексных HTML-страниц:

Конструкция	описание	Комментарии
<~rubric start~>	Маркер начала рубрики	Все содержимое между этими маркерами будет повторяться столько раз, сколько имеется рубрик в системе. Причем внутри каждого такого повторяющегося блока информация будет извлекаться только из соответствующей экземпляру блока рубрики.
<~rubric stop~>	Маркер конца рубрики	
<~rubric name~>	Название текущей рубрики	Конструкция заменяется на название текущей рубрики. Актуально только между маркерами начала и конца рубрики.
<~news start~>	маркер начала заголовка новости	Все содержимое между этими маркерами будет повторяться для каждого сообщения (новости) из списка экспортируемых. Внутри маркеров значения следующих переменных берутся для конкретных сообщений.
<~news stop~>	маркер конца заголовка новости	
<~news date~>	дата новости	Заменяется на дату конкретной новости. Актуально только внутри маркеров начала и конца заголовка новости.
<~news time~>	время новости	Заменяется на время конкретной

		новости. Актуально только внутри маркеров начала и конца заголовка новости.
<~news link~>	Ссылка на новость	Заменяется на ссылку на новостную страницу, создаваемую системой в процессе экспорта. Актуально только внутри маркеров начала и конца заголовка новости.
<~news url~>	url страницы-источника	Заменяется на ссылку на страницу-источник в Интернете.
<~news title~>	заголовок новости	Заменяется на заголовок конкретной новости. Актуально только внутри маркеров начала и конца заголовка новости.
<~news subtitle~>	подзаголовок новости	Заменяется на подзаголовок конкретной новости. Актуально только внутри маркеров начала и конца заголовка новости.
<~news text~>	текст новости	Заменяется на текст новости.
<~current date~>	текущая дата	Заменяется на текущую дату (дату экспорта)
<~current time~>	текущее время	Заменяется на текущее время (время экспорта)
<~change {фрагмент1}{фрагмент2} change~>	чередующиеся с новостями фрагменты HTML-текста	Используется внутри маркеров заголовка новости. Для каждой следующей новости заменяется на чередующееся значение, указанное в конструкции. Например, для первой новости вместо конструкции будет подставлен <i>фрагмент1</i> , для второй – <i>фрагмент2</i> , для третьей – <i>фрагмент1</i> и т.д.

Предопределенные конструкции в шаблонах новостных HTML-страниц:

Конструкция	описание	Комментарии
<~news title~>	заголовок новости	Заменяется на заголовок новости.
<~news text~>	текст новости	Заменяется на текст новости.
<~news date~>	дата новости	Заменяется на дату новости.
<~news time~>	время новости	Заменяется на время новости.
<~news url~>	URL новости	Заменяется на ссылку на страницу-источник в Интернете.
<~news site~>	Название источника новости	Заменяется на название источника новости.

<~news subtitle~>	подзаголовок новости	Заменяется на подзаголовок новости.
<~current date~>	текущая дата	Заменяется на текущую дату (дату экспорта)
<~current time~>	текущее время	Заменяется на текущее время (время экспорта)

Конструкция для пользовательской переменной имеет вид:

<~имя_переменной~> - при экспорте пользователем будет указываться значение этой переменной, которое подставляется вместо конструкции.

Пример HTML-шаблона индексной страницы. Объявлена пользовательская переменная *pagetitle*:

```
<HTML>
<HEAD>
<TITLE> <~pagetitle~> </TITLE>
<META http-equiv=Content-Type content="text/html; charset=windows-1251">
<BODY>
<h1> <~pagetitle~> </h1>
<~rubric start~>
  <h2> <~rubric name~> </h2>
  <~news start~>
    <~news date~> <~news time~>
    <a href=<~news link~>><~news title~></a><br>
    <~news subtitle~> <br>
  <~news stop~>
  <br>
<~rubric stop~>
<br>
Страница создана программой Avalanche Lite <~current date~> <~current time~>

</body>
</html>
```


Настройка экспорта в формате XML

Система экспорта не ограничивает тип выходных документов форматом HTML в рамках текстового типа файлов. То есть выходные документы могут быть, к примеру, XML-документами, или SQL – выражениями. В связи с удобством и популярностью формата XML приведем пример настройки шаблонов для экспорта сообщений в XML.

Пример XML шаблона индексной страницы:

```
<?xml version="1.0" encoding="windows-1251"?>
<exportset name="<~pagetitle~>" date="<~current date~>" time="<~current time~>">
<~rubric start~>
  <rubric name="<~rubric name~>">
    <~news start~>
      <news date="<~news date~>" time="<~news time~>">
        <link> <~news link~> </link>
        <title> <~news title~> </title>
        <subtitle> <![CDATA[ <~news subtitle~> ]]> </subtitle>
      </news>
    <~news stop~>
  </rubric>
<~rubric stop~>
</exportset>
```

Пример XML-шаблона новостной страницы:

```
<?xml version="1.0" encoding="windows-1251"?>
<exportnews date="<~current date~>" time="<~current time~>">
  <title> <~news title~> </title>
  <date> <~news date~> </date>
  <time> <~news time~> </time>
  <newstext> <![CDATA[<~news text~>]]> </newstext>
  <url> <~news url~> </url>
  <site> <~news site~> </site>
</exportnews>
```

Просмотр статистики поступления

Статистика позволяет наблюдать график поступления сообщений в определенную рубрику с течением времени.

Для получения доступа к статистике:

1. Выберите рубрику в списке рубрик
2. нажмите правую кнопку мыши на выбранной рубрике и выберите пункт меню "Показать статистику". Появится окно с графиком, отображающее количество документов по оси ординат (Y) и время в днях (7 дней) по оси абсцисс (X):



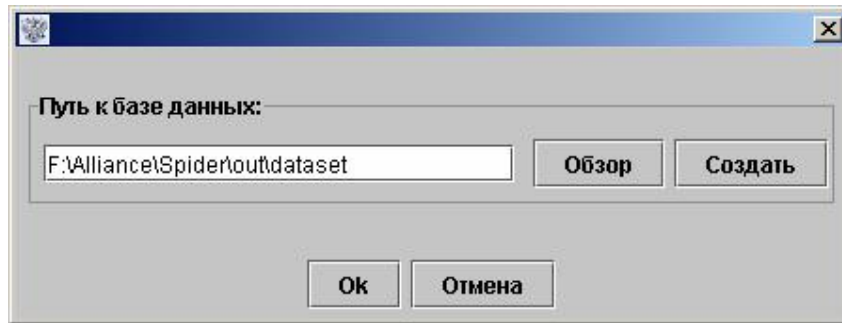
Настройка базы данных

Настройка базы данных заключается в двух операциях:

- создание базы данных
- изменение пути к базе данных

Вместе с программой поставляется уже созданная база данных, таким образом использовать настройки следует только в случае, если вы хотите разместить файлы базы данных в другом месте (например на другом носителе), использовать базу данных, оставшуюся после удаления и переустановки программы, или оперировать несколькими БД.

Доступ к настройкам можно получить через меню *Файл->Поменять рабочую базу программы просмотра*:



Для создания базы данных:

1. Введите путь к папке, где будут храниться файлы БД.
2. Нажмите "Создать".

Для изменения пути к БД:

1. Введите путь, либо используйте кнопку "Обзор" и выберите любой файл БД в папке, которая ее содержит.
2. Нажмите ОК.

База данных содержит следующие файлы:

- ava.jds
- ava_LOGA_*
- ava_STATUS_*
- ava_temp.jds
- datahub.dat

Для сохранения или перемещения базы данных (т.е. всех сообщений и настроек) достаточно скопировать или переместить все вышеперечисленные файлы в другую директорию.

Настройка прокси-сервера.

В случае, если работ в сети происходит посредством прокси-сервера, необходимо указать настройки прокси сервера для использования последних программным средством Avalanche. Для настройки прокси-сервера необходимо: в меню «Управление» поискового агента Spider необходимо выбрать пункт «Настройка прокси». Затем, необходимо указать собственно настройки прокси-сервера в появившемся диалоговом окне. Нажатие кнопки «По умолчанию» приводит к получению настроек прокси-сервера из реестра системы. В этом случае используются настройки прокси-сервера, указанные для Internet Explorer.

Настройка прокси

Настройка прокси

Host: 192.123.132.12

Port: 8081

Применить По умолчанию

Проверка

http://www.ya.ru Go

Работа с поисковым агентом (Spider)

Материал раздела содержит инструкции по работе с программой - поисковым агентом, основными задачами которой является сбор сообщений в сети Интернет, классификация их по рубрикам и сохранение в базе данных для дальнейшей работы.

Если вы только начинаете работу с программой, рекомендуется посмотреть описание процесса функционирования системы, а также минимальную информацию, необходимую для начала работы с системой, чтобы приступить к работе прямо сейчас.

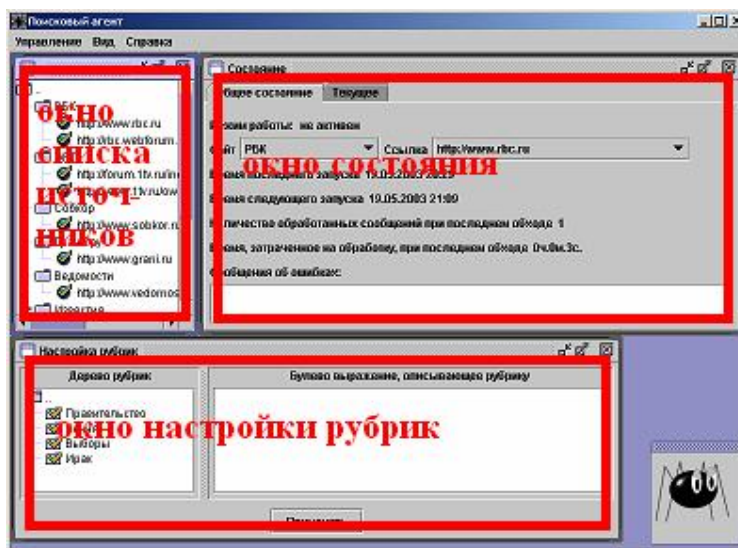
Для опытных пользователей будет интересно прочитать подробное руководство по настройке системы.

Для удобства работы с программой приведен список горячих клавиш.

Функционирование программы

В главном окне системы отображаются:

- окно списка источников, предназначенное для размещения списка сайтов и страниц, с которых собираются сообщения;
- окно отображения состояния программы;
- окно настройки тематических рубрик;
- главное меню программы
- паучок,двигающийся во время сбора сообщений в Сети.



Рабочий цикл программы состоит из следующих операций:

- настройка системы: формирование списка сайтов, настройка рубрик, указание политики сбора сообщений и каталога на машине пользователя, в который будут помещаться файлы с сообщениями.
- запуск системы в одном из режимов (ручной, таймер);
- контроль за ходом выполнения программы;
- просмотр, поиск, экспорт, новостей, а также просмотр статистики.

Главное меню

Меню Управление - координация основных функций программы

Старт	- команда начала сбора сообщений в ручном режиме
Стоп	- команда насильственного прекращения сбора.
Активировать таймер	- переход в режим работы системы сбора сообщений по таймеру
Завершить работу таймера	- переход в режим только ручного сбора сообщений без таймера
Настройка	- диалог настройки базы данных и политики сбора
Настройка прокси	- диалог настройки работы через прокси сервер
Выход	- выход из программы

Меню Вид - настройки внешнего вида и отображаемых окон

Показать список сайтов	- отображение окна со списком сайтов
Показать состояние	- отображение окна состояния.
Показать окно рубрик	- отображение окна с настройками рубрик
Показать лог-файл	- просмотр файла, в который записываются действия программы во время сбора сообщений
Просмотр результатов	- вызов программы-поисковика Avalanche

Меню **Справка** - руководство пользователя и информация о программе

Справка по системе - вызов данного руководства пользователя

О программе... - информация о программе и ее создателях

Рабочий цикл программы

Рабочий цикл программы-паука Spider состоит из следующих операций:

- настройка системы (при первом использовании можно пропустить):
 - формирование списка сайтов,
 - настройка рубрик,
 - указание политики сбора сообщений и
 - настройка базы данных.
- запуск системы в одном из режимов (ручной, таймер);
- контроль за ходом выполнения программы ;
- просмотр и поиск новостей.

Настройка системы

Самым сложным и самым трудоемким этапом работы с системой является этап настройки. В процессе эксплуатации системы также неизбежно периодическое обращение к системе настройки. Связано это с тем, что сайты со временем могут вводить новые разделы, менять дизайн, форму подачи материала и т.д. Настройка системы заключается в следующем:

- формировании списка сайтов - создание списка сайтов с новостными лентами или сайтов-поисковиков, с которых будут собираться сообщения;
- настройке рубрик - рубрики фильтруют сообщения согласно заданным условиям, например, по ключевым словам или источникам.
- определении политики сбора - политика сбора определяет необходимо ли Вам сохранять все сообщения при сборе, или только те, которые принадлежат рубрикам.
- настройке базы данных - база данных - это хранилище ваших настроек и сообщений (новостей). Вы можете расположить ее в удобном Вам месте, или иметь несколько баз данных.

Формирование списка сайтов

Формирование поискового пространства осуществляется исходя из целей и задач поиска, стоящих перед пользователем. Формирование поискового пространства включает задание списка новостных сайтов и сайтов поисковых систем, с которых будет собираться информация.

Для сбора новостных сообщений с сайтов, содержащих новостные ленты необходимо:

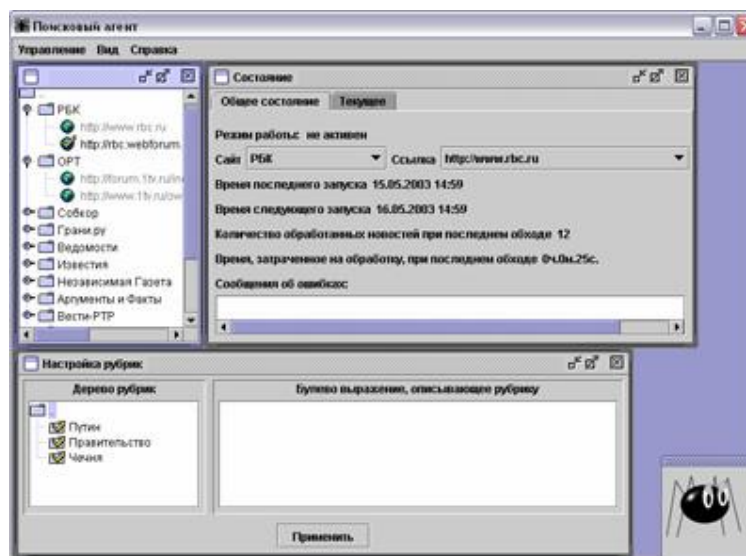
1. Задать список сайтов со страницами, адекватными информационным потребностям пользователя;
2. Произвести настройку этих сайтов, определяющую ссылки в Сети (URL) на страницы, выбранные пользователем, и процедуры их обработки.

Для сбора результатов поиска поисковых систем Интернета

1. Необходимо создать список поисковых систем в программе и задать настройки для каждой из них;
2. Добавить для каждой поисковой системы набор запросов на поиск.

Работа со списком сайтов

В исходном состоянии программа содержит список из нескольких известных сайтов. При эксплуатации системы задание поискового пространства для конкретной задачи сбора документов в сети Интернет сводится к редактированию имеющихся списка сайтов и списка ссылок в соответствии с потребностями пользователя.



Для работы со списками сайтов и ссылок используется левое верхнее окно (см рис. выше), если оно закрыто, то необходимо выбрать **Вид** и затем **Показать список сайтов** (откроется окно с деревом сайтов/ссылок - левая панель окна

«Поисковый агент» на приведенном выше рисунке)

Добавление сайта в список сайтов

1. Кликнуть правой кнопкой на корне дерева (самом верхнем элементе с меткой "..") и выбрать пункт *«Добавить сайт»*;
2. Ввести название сайта и нажать **ОК** ;
3. Для добавления ссылок к этому сайту см **«Добавление ссылок»**.

Добавление поисковой системы

1. Щелкнуть правой кнопкой на корне дерева ("..") и выбрать пункт "Добавить поисковик";
2. Ввести название поисковой системы, начало ссылки (строку до текста запроса), конец ссылки (строку после текста запроса) и указать стоит ли формировать запросы по существующим рубрикам. Более подробную информацию о настройке новостей смотрите здесь.
3. Нажмите ОК.

Удаление сайта или поисковой системы

1. Выбрать сайт, который необходимо удалить.
2. Кликнуть на нем правой кнопкой мыши
3. Выбрать пункт "Удалить сайт"
4. Нажать Ок.

Переименование сайта

1. Выбрать сайт.
2. Кликнуть правой кнопкой мыши
3. Выбрать пункт «Изменить название сайта»
4. Ввести новое название сайта
5. Нажать ОК.

Добавление ссылок

1. Выбрать сайт, к которому надо добавить ссылку.
2. Кликнуть правой кнопкой мыши на названии этого сайта и выбрать пункт *Добавить ссылку*.
3. В появившемся окне ввести ссылку на ресурс в Сети (URL) с указанием протокола (<http://>) и нажать ОК.

Удаление ссылки

1. Выбрать ссылку, которую необходимо удалить.
2. Кликнуть на ней правой кнопкой мыши
3. Выбрать пункт «Удалить ссылку»
4. Нажать Ok.

Просмотр состояния ссылки

1. Выбрать ссылку.
2. Кликнуть правой кнопкой мыши
3. Выбрать пункт «Состояние»
4. В окне «Состояние» отобразится касающаяся этой ссылки информация, а именно когда обошел ее поисковый агент, сколько новостей с нее было собрано,

Временное исключение из поискового пространства.

1. Выбрать ссылку.
2. Кликнуть правой кнопкой мыши
3. Выбрать пункт «Обрабатывать/Не обрабатывать»
4. Цвет ссылки изменится с черного на серый, картинка измениться (пропадет "галочка"), и ссылка не будет обрабатываться поисковым агентом. Для того чтобы включить ссылку в поисковое пространство снова, необходимо проделать ту же операцию. Аналогичного эффекта можно достичь двойным кликом по ссылке.

Настройка ссылки

1. Выбрать ссылку.
2. Кликнуть ссылку правой кнопкой мыши
3. Выбрать пункт «Свойства»
4. Отобразится окно свойств, в котором производятся настройки.

Настройка ссылок

Для настройки ссылки необходимо

1. Выбрать ссылку.
2. Кликнуть ссылку правой кнопкой мыши

3. Выбрать пункт «Свойства»

Отобразится окно свойств, изображенное ниже

The screenshot shows a dialog box titled 'http://www.strana.ru/news/'. It contains the following elements:

- Address bar: `http://www.strana.ru/news/`
- Checkboxes:
 - Временно исключить из списка
 - Собирать только новости с новостных лент
 - Не производить чистку документа
- Input field: Периодичность/время обхода: 1440
- Checkbox: Дата сообщения находится после заголовка
- Text boxes:
 - Обрабатывать ссылки, находящиеся в документе после: []
 - Обрабатывать ссылки, находящиеся в документе перед: []
- Text boxes:
 - Обрабатывать только ссылки, имеющие следующий вид: []
 - НЕ обрабатывать ссылки, имеющие следующий вид: []
- Dropdown: Формат даты: []
- Text boxes:
 - Ссылки на следующие страницы: []
 - Глубина: []
- Code editor: Скрипт дополнительной настройки: `<div class=brclear><br clear=all></div>`
- Text box: Принадлежность рубрикам: []
- Buttons: Редактировать (twice), Применить

В открывшемся диалоговом окне содержатся адрес (URL) ссылки, бокс для отметки **<Временно исключить из списка>** и другие поля для ввода атрибутов выделенной ссылки. Настройка ссылки заключается в заполнении всех или некоторых полей этого диалога и нажатия кнопки «Применить».

- Для изменения адреса ссылки необходимо изменить надпись в поле «Адрес»
- "**Временно исключить из списка**" текущую ссылку можно поставив отметку здесь, или сделав двойной щелчок на в окне списка источников.
- В поле "**Периодичность/время обхода**" указывается периодичность, с которой система будет забирать новости с указанного адреса. Периодичность задается в минутах. По умолчанию задаётся число 1440, что значит заходить на страницу каждые 1440 минут, т.е. один раз в сутки. Кроме того в этом поле можно указать непосредственно время обхода, например 12:00, в этом случае при работе таймера каждый день в 12:00 будет обрабатываться эта ссылка.
- Отметку "**Собирать только новости с новостных лент**" необходимо использовать когда текст новости находится прямо на ленте, и нет необходимости ходить по ссылкам.
- Как правило, на новостных лентах сначала указывается дата и время, а потом заголовок. В случае, когда это не так: вначале идет заголовок, потом анонс новости, а только после этого дата и время, необходимо поставить «галочку» в поле "**Дата сообщения находится после заголовка**".

- Отметку **"Не производить чистку документов"** в большинстве случаев нужно ставить если используется скрипт дополнительной настройки или особый формат ссылки. Иначе предварительная очистка HTML от незначимых тегов и конструкций приведет к потере информации и неправильному определению новости.
- **"Обрабатывать ссылки, находящиеся в документа после"** указанной строки означает, что весь html-текст до первого вхождения указанного фрагмента будет игнорироваться.
- **"Обрабатывать ссылки, находящиеся в документе до"** введенной строки также игнорирует все, после первого вхождения этого фрагмента, с учетом предыдущей настройки.
- В поле **"Обрабатываются только ссылки..."** нужно ввести строку-шаблон вида: <значение>*<значение>*<значение>... Ссылки другого вида обрабатываться не будут. Например будет обрабатывать ссылки только указанного формата, где вместо символа "звездочка" находится любая строка.
- Поле **"Не обрабатывать ссылки..."** работает аналогично, но игнорирует все ссылки, удовлетворяющие шаблону. Шаблон имеет точно такой же формат, что и в предыдущей настройке.
- Определяется **формат даты**, используемый на сайте новостей.
- Строка **"Ссылки на следующие страницы"** задает формат формат ссылок, на которых находится продолжение новости.
- Поле **"Глубина"** указывает на сколько ссылок на следующие страницы можно перейти за один обход.
- Поле **"Скрипт дополнительной настройки"** в диалоговом окне атрибутов ссылки предназначено для более гибкой настройки, в частности для того, чтобы собирать новости с аналитических сайтов, необходимо написать инструкцию "\$All". Подробнее о дополнительной настройке. Обычно, при использовании дополнительной настройки необходимо поставить отметку **"Не производить чистку документов"**.
- Достаточно часто новостные и аналитические сайты публикуют свои материалы в нескольких различных разделах. Каждый раздел представляет собой совокупность сообщений по определенной тематике. Например: Лента «В России», «В мире», «В регионах», «Политика», «Экономика» и т.д. Некоторые ленты по смыслу могут соответствовать рубрикам системы. В этом случае необходимо воспользоваться полем **«Принадлежность рубрикам»**, которое служит для того, чтобы принудительно все новости из данного раздела помещать в указанную рубрику. Возможно вы захотите определить для этого дополнительные рубрики.

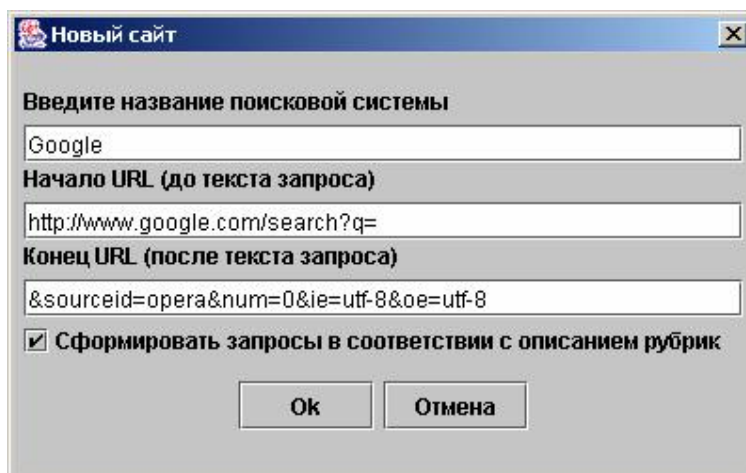
Настройка поисковых систем

Наравне со сбором новостей с информационных сайтов, существует возможность делать запросы по ключевым словам к существующим поисковым системам Интернета, таким как Яндекс (www.yandex.ru), Rambler (www.rambler.ru), Google (www.google.ru) и многим другим. В роли сообщения в данном случае выступает один элемент(ссылка) результата поиска в такой системе.

Настройка нового поисковика разделяется на два этапа:

1. Создание сайта

- вызвать правой кнопкой мыши контекстное меню корневого элемента ("..") списка сайтов и выбрать пункт "добавить поисковик". Появится диалоговое окно, показанное на рисунке:



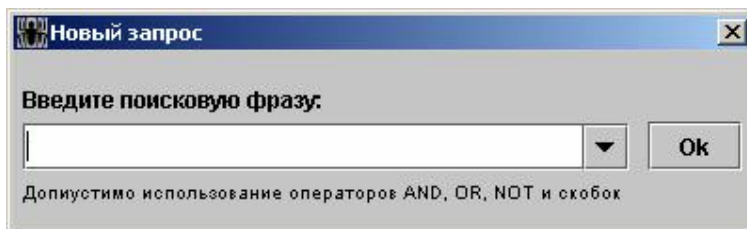
- В окне требуется ввести название поисковой системы, которое будет отображаться в списке сайтов, часть ссылки до текста запроса и часть ссылки, которая следует после текста запроса.

Примечание: Обычно логика работы поисковых систем следующая: вы вводите в браузере адрес поисковика (например, www.google.com), затем на главной странице в окне редактирования печатаете ключевые слова и нажимаете кнопку "поиск". При этом браузер открывает страницу с результатами поиска, ссылка на которую содержит в качестве параметра ваш запрос. Таким образом, ссылка имеет следующий вид: <строка до запроса><ваши ключевые слова><строка после запроса> (без угловых скобок). Например:
<http://www.google.com/search?q=Hello&sourceid=opera&num=0&ie=utf-8&oe=utf-8>

Установите отметку "Сформировать запросы в соответствии с описанием рубрик", если хотите чтобы после нажатия кнопки ОК у вас автоматически сформировались запросы к поисковой системе по существующим рубрикам.

2. После внесения нового поисковика в список сайтов, можно создать набор запросов на поиск к этим поисковикам. Если на предыдущем этапе вы отметили "Сформировать запросы в соответствии с описанием рубрик", то у вас уже есть столько запросов, сколько рубрик было во время внесения поисковика.

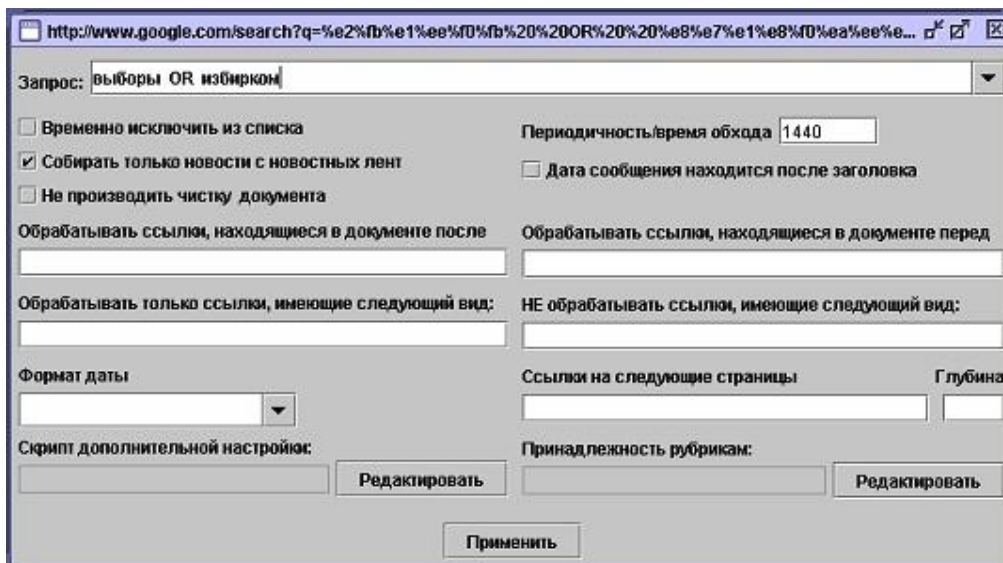
Для добавления нового запроса щелкните правой кнопкой мыши по поисковикам в списке сайтов и выберите пункт **"Добавить ссылку"**:



Здесь введите набор ключевых слов, по которым будет производиться поиск, возможно используя логические операторы AND (конъюнкция - И), OR (дизъюнкция - ИЛИ) и NOT (отрицание - не). Затем нажмите ОК.

3. Настройка ссылки.

Для открытия диалога настройки ссылки-запроса к поисковой системе нажмите "Свойства" в контекстном меню ссылки в дереве списка сайтов.



Настройка ссылки-запроса во многом аналогична настройке обычной новостной ссылки, однако имеются также и отличия.

- Для изменения текста запроса необходимо изменить текст в поле **«Запрос»**
- **"Временно исключить из списка"** текущий запрос можно поставив отметку здесь, или сделав двойной щелчок на в окне списка сайтов.

- В поле "**Периодичность/время обхода**" указывается периодичность, с которой система будет проверять результаты поиска в поисковике. Сохраняться, конечно, будут только новые результаты. Периодичность задается в минутах. По умолчанию задаётся число 1440, что значит заходить на страницу каждые 1440 минут, т.е. один раз в сутки. Кроме того в этом поле можно указать непосредственно время обхода, например 12:00, в этом случае при работе таймера каждый день в 12:00 будет обрабатываться эта ссылка.
- Отметку "**Собирать только новости с новостных лент**" в данном случае необходимо использовать всегда.
- Поле "**Дата сообщения находится после заголовка**" в данном случае не имеет значения.
- Отметку "**Не производить чистку документов**" в большинстве случаев нужно ставить если используется скрипт дополнительной настройки или особый формат ссылки. Иначе предварительная очистка HTML от незначимых тегов и конструкций приведет к потере информации и неправильному определению новости.
- "**Обрабатывать ссылки, находящиеся в документе после**" указанной строки означает, что весь html-текст до первого вхождения указанного фрагмента будет игнорироваться.
- "**Обрабатывать ссылки, находящиеся в документе до**" введенной строки также игнорирует все, после первого вхождения этого фрагмента, с учетом предыдущей настройки.
- В поле "**Обрабатываются только ссылки...**" нужно ввести строку-шаблон вида: <значение>*<значение>*<значение>... Ссылки другого вида обрабатываться не будут. Например будет обрабатывать ссылки только указанного формата, где вместо символа "звездочка" находится любая строка.
- Поле "**Не обрабатывать ссылки...**" работает аналогично, но игнорирует все ссылки, удовлетворяющие шаблону. Шаблон имеет точно такой же формат, что и в предыдущей настройке.
- Значение **формата даты** не влияет на работу поисковика.
- Строка "**Ссылки на следующие страницы**" в данном случае задает формат формат ссылок, на которых находится продолжение результатов поиска, так как обычно поисковик дает лишь по 10-20 результатов на страницу.
- Поле "**Глубина**" указывает на сколько ссылок на следующие страницы можно перейти за один обход. Фактически, этот параметр указывает какое максимальное количество результатов поиска рассматривать. Если в поисковике на страницу выводится 10 результатов, то значение этого поля равное 3 сохранит максимум 30 результатов.

- Поле "**Скрипт дополнительной настройки**" в диалоговом окне атрибутов ссылки предназначено для более гибкой настройки, в частности для указания шаблона страницы поисковика. Подробнее о дополнительной настройке. Обычно, при использовании дополнительной настройки необходимо поставить отметку "**Не производить чистку документов**".
- Если настраиваемая поисковая система имеет четко определенную тематику результатов (например, поиск по книжным новинкам) , то можно воспользоваться полем «**Принадлежность рубрикам**», которое служит для того, чтобы принудительно все результаты поиска помещать в указанную рубрику. Возможно вы захотите определить для этого дополнительные рубрики.

Тонкая настройка

"*Дополнительная*" или "*Тонкая*" настройка предоставляет достаточно широкие возможности для настройки системы для работы практически с любым сайтом. Достаточно часто требуется хорошее знание HTML.

В диалоговом окне необходимо кликнуть мышью на кнопке «Редактировать» напротив поля "Скрипт дополнительной настройки". Откроется редактор, в окне которого необходимо написать скрипт в соответствии с нижеследующим синтаксисом. По окончании написания скрипта необходимо нажать кнопку «Сохранить» и закрыть окно редактора.

Скрипт состоит из блоков вида:

<параметр><пробел><значение>

Типы блоков:

\$GetLinksLongerThen <значение> - определяется минимальная длина ссылки (количество видимых символов). Например,

\$GetLinksLongerThen 10 значит, что обрабатываются только те ссылки, количество символов в которых больше 10.

мама мыла не будет обработана («мама мыла» -9 символов), а

мама мыла раму -будет.

Для точной настройки системы под конкретный сайт можно определить три вида шаблонов:

1. Шаблон новостей используется для разбора первой новостной страницы, содержащей ленту новостей.

2. Шаблон текста новости используется для разбора страницы, на которой находится полный текст одной новости. Предполагается что на текстовые

страницы ведут ссылки с новостной страницы. Формат этих ссылок также можно указать в основных настройках.

3. Шаблон сообщений форума определяет структуру сообщений на странице, где пользователи оставляют отзывы о новости.

Формат шаблона новостной ленты

\$NewsShablon <значение> - позволяет полностью определять новость.

Пример:

```
$NewsShablon <a name="~something~"></a><table border=0 cellpadding=0
cellspacing=5>
<tr valign=top><td width=15><font color="green"><b>~time~</b></font></td>
<td width=1></td>
<td width=615><p><b>~title~</b></p><p><b>~text~</b></p>
</tr><tr valign="top"><td colspan="2">&nbsp;</td><td align="right">
<table cellspacing=0 cellpadding=0 border=0 width="100%"><tr valign="top">
<td align="left" width="50%">~something~</td><td align="right"
width="50%">~something~</td></tr></table></td></tr></table>
```

HTML-код, описывающий несколько новостей на одной странице, практически идентичен, поэтому в HTML-коде необходимо выявить повторяющиеся куски, скопировать его и заменить изменяющиеся части на соответствующие параметры:

~something~ - меняющийся, но интересующий нас текст.

~title~ - текст, стоящий на этом месте, является заголовком новости.

~subtitle~ - текст, стоящий на этом месте, является анонсом новости.

~date~ - текст, стоящий на этом месте, является датой новости.

~time~ - текст, стоящий на этом месте, является временем новости.

~url~ - текст, стоящий на этом месте, является адресом страницы, на которой находится текст новости.

~text~ - текст, стоящий на этом месте, является текстом новости.

~source~ - источник сообщения.

~autor~ - автор сообщения.

Формат текстового шаблона

\$TextShablon <значение> - позволяет полностью определять вид документа, содержащего текст сообщения. Параметры те же, что и в новостном шаблоне кроме `~url~` и `~subtitle~` (их нет)

Формат шаблона сообщений форума

\$TalkShablon <значение> - аналогичен **\$TextShablon**, за тем исключением, что ищет повторяющиеся куски информации. Используется только для определения отзывов на статью и сообщений в форуме (когда их несколько на одной странице). Еще одним отличием является наличие двух вспомогательных параметров, которые пишутся в начале:

`~~StartsWith` <значение>~~ - означает, что поиск повторяющихся сообщений необходимо искать с <значение>

`~~EndsWith` <значение>~~ - означает, что закончить поиск повторяющихся сообщений необходимо с достижением <значение>

Пример:

```
$TalkShablon ~StartsWith Отзывы посетителей~ ~EndsWith
</td></tr></table>~ <b>~author~</b> &nbsp;~time~<br>~text~<p>
```

\$# - комментарии.

Внимание! При использовании скрипта рекомендуется отключать чистку!

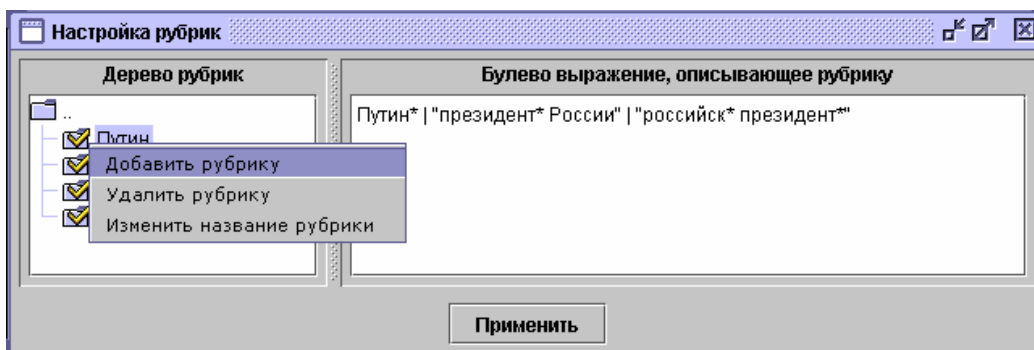
Настройка рубрик

Рубрикой называется некоторое множество документов, принадлежащих определенной тематике, в узком или широком смысле этого слова. В данной системе рубрика задается путем описания ее смысла средствами булевой алгебры, то есть множеством ключевых слов и словосочетаний, объединенных логическими операциями "И", "ИЛИ", "НЕ".

Рубрики занимают важнейшее место в логической архитектуре системы. Хорошо настроенные рубрики позволят Вам сэкономить много времени при поиске интересующих вас сообщений и помогут классифицировать поступающие сообщения для создания тематического архива новостей, которые в дальнейшем можно экспортировать в HTML-файлы и опубликовать в Сети. Поэтому очень важно предварительно качественно настроить фильтры рубрик, иначе можно не справиться с потоком поступающих сообщений.

Подробнее об настройке рубрик речь пойдет в разделе Настройка булева фильтра. Однако перед этим стоит прочитать механизм управления списком рубрик (добавление, удаление, переименование и временное исключение).

Работа с перечнем рубрик



Для работы с перечнем рубрик в системе предусмотрено окно "Настройка рубрик". Состоит оно из двух частей: левое для отображения перечня и структуры рубрик, правое - для описания (указания) смысла рубрики.

- Для добавления новой рубрики необходимо кликнуть на корне дерева или на существующей рубрике правой клавишей мыши. Из появившегося меню (см. рис.) выбрать пункт Добавить рубрику. В появившемся диалоговом окне ввести название, нажать кнопку "Ок", рубрика будет добавлена (в первом случае в качестве рубрики, во втором в качестве подрубрики той рубрики, по которой кликнули).
- Удаляется рубрика путем щелчка правой клавиши мыши на удаляемой рубрике и выбора пункта "Удалить рубрику".
- Название рубрики изменяется аналогично выбором соответствующего пункта из контекстного меню.
- Для описания рубрики необходимо выделить ее, в правом окне задать булев фильтр и нажать кнопку применить. О синтаксисе описания рубрик подробно написано в следующем разделе.

Настройка булева фильтра

В данной системе используется рубрикация, основанная на булевой алгебре. В основе этого метода лежит вычисление булева выражения, определяющего рубрику, типа

(Собака & кошка) / мышка

В этом выражении все слова заменяются на "истина" или "ложь" в зависимости от того, есть ли это слово в документе или нет, после чего вычисляется выражение. Если значение выражения "истина" - документ принадлежит рубрике, если "ложно" - нет.

Описанный выше пример является классическим вариантом. В данной системе реализован достаточно мощный и гибкий способ задания рубрик, позволяющий:

- использовать логические операторы:
 & - логическое И
 | - логическое ИЛИ
 ! - логическое НЕ
- использовать скобки ()
- задавать словоформы - используя символ * можно задавать шаблоны слов, например: собак*. В этом случае будут искажаться слова, начинающиеся с сочетания букв собак.
- задавать словосочетания: достаточно часто необходимо чтобы искомые слова стояли рядом друг с другом. Например, "президент России", только при условии, что эти слова идут рядом и друг за другом можно сказать, что в документе речь идет про Путина. В данной системе это реализовано посредством использования двойных кавычек. ("президент* России")
- задавать названия и имена собственные. Если слово в фильтре указано с большой буквы, то меняется оно на "истина" лишь в том случае, если в тексте это слово также написано с большой буквы.
- реализован механизм поиска слов на определенном расстоянии друг от друга.
Синтаксис: [("словосоч" | "словосоч*"|...) ("словосоч*" | "словосоч*"|...)]N*
 если одно из словосочетаний в первой скобке стоит не далее чем на расстоянии N слов от одного из словосочетаний из второй скобки, то значение выражения - истина. Вместо круглых скобок может стоять одно слово или словосочетание. Например, если необходимо найти документ в котором слово драка стоит не далее пяти слов от фамилии Жириновский необходимо написать следующее
 [Жириновск* драк*]5 или [(Жириновск* | "Владимир* Вольфович*") драк*]5

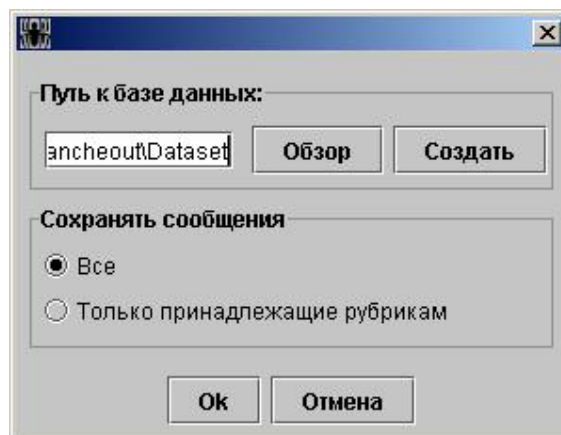
Пример:

Ищем все, что касается драк Жириновского в Думе

[(Жириновск* | "Владимир* Вольфович*") (драк* | подрал*)]15 & дум*

Определение политики сбора

Для определения политики сбора, необходимо выбрать в меню **Управление->Настройка**.



В появившемся окне указать политику сбора. Политика сбора - это один из режимов сбора сообщений:

1. собирать все сообщения с указанных ресурсов и сохранять их на машине пользователя.
2. собирать только сообщения, принадлежащие рубрикам. В этом режиме в базе данных сохраняются лишь те сообщения, которые принадлежат хотя бы одной рубрике.

Для выбора первого режима необходимо выбрать пункт "Все", для выбора второго - "Только принадлежащие рубрикам".

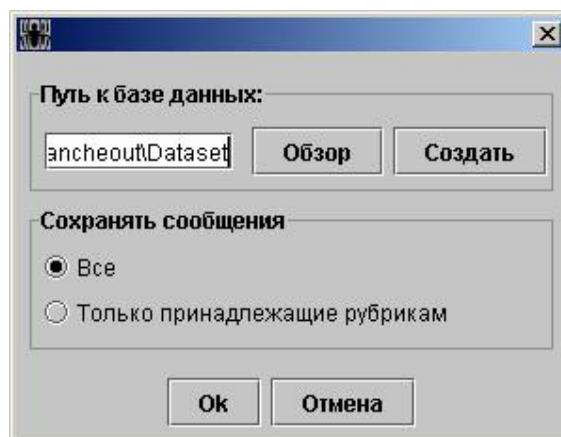
Настройка базы данных

Настройка базы данных заключается в двух операциях:

- создание базы данных
- изменение пути к базе данных

Вместе с программой поставляется уже созданная база данных, таким образом использовать настройки следует только в случае, если вы хотите разместить файлы базы данных в другом месте (например на другом носителе), использовать базу данных, оставшуюся после удаления и переустановки программы, или оперировать несколькими БД.

Доступ к настройкам можно получить через меню *Управление->Настройка* поискового агента Spider:



Для создания базы данных введите путь к папке, где будут храниться файлы БД и нажмите "Создать".

Для изменения пути к БД введите путь, либо используйте кнопку "Обзор" и выберите любой файл БД в папке, которая ее содержит.

База данных состоит из следующих файлов:

- ava.jds
- ava_LOGA_*
- ava_STATUS_*
- ava_temp.jds
- datahub.dat

Сохранение настроек

Настройки системы хранятся в файлах, расположенных в корневом каталоге программы. Это:

doplinkinfo.dat - содержит статистическую информацию по ссылкам: сколько сообщений было собрано в последний раз, за какое время ...

spidersettings.dat - содержит информацию о текущих настройках, например путь к базе данных.

файлы, имена которых начинаются с "http" - содержат информацию о обработанных ссылках при последнем обходе

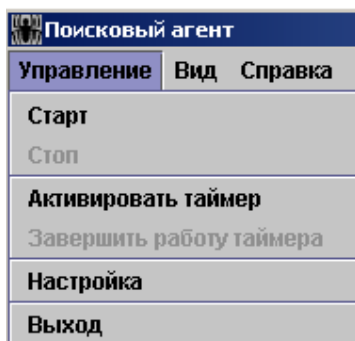
остальные настройки, а также **все собранные сообщения** находятся в базе данных (по умолчанию в каталоге Dataset).

Для переноса настроек от одной копии системы к другой достаточно скопировать эти файлы, хотя в большинстве случаев можно ограничиться лишь копированием **базы данных**.

Сбор сообщений

Для поиска и сбора сообщений в системе предусмотрены два режима работы: **Ручной** и **Таймер**. Система может работать только в одном из них. Выбор режима осуществляется из меню "**Управление**".

Выберите "**Старт**" для ручного управления и "**Активировать таймер**" для режима "**Таймер**".



В ручном режиме программа просматривает список сайтов, начиная сверху, и ставит в очередь все их ссылки. Паук обрабатывает ссылки в порядке постановки в очередь. При этом время последнего запуска запоминается. Данный режим завершает свою работу, когда обработаны все ссылки на всех сайтах, или когда пользователь выбирает пункт "Стоп" в меню "Управление".

В режиме работы по таймеру система для каждой ссылки постоянно проверяет, не превысило ли время с момента ее последнего запуска некоторое значение, установленное настройкой. Если время превышено, то программа начинает обрабатывать данную ссылку. Одновременно могут обрабатываться четыре ссылки. Если при этом наступает время обработки пятой, то она ставится в очередь. Работает этот режим до тех пор, пока не будет выключен пользователем путем выбора пункта "Завершить работу таймера" в меню "Управление", или не будет закрыта программа.

Стоит заметить, что программа не может одновременно работать и в том, и в другом режиме. Для переключения между режимами необходимо завершить текущий и после включить требуемый.

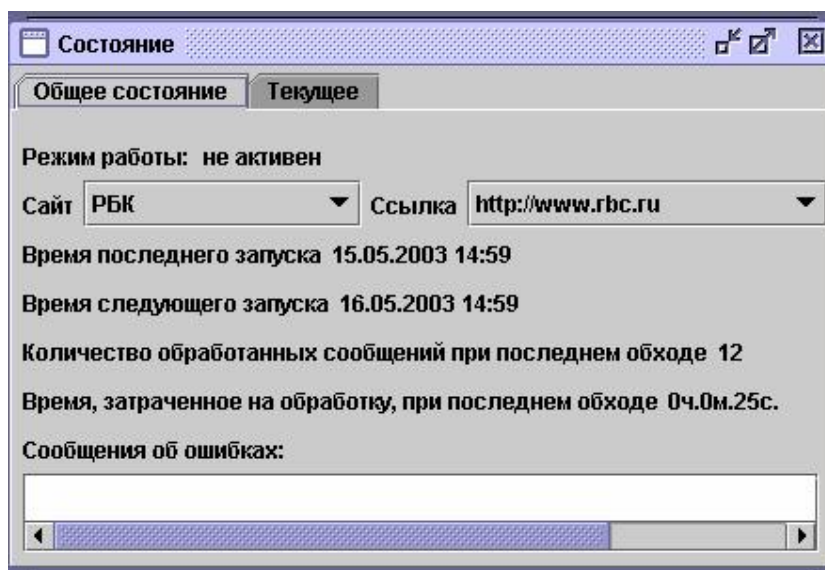
Внимание! В обоих случаях программа запоминает адреса страниц обработанных при последнем обходе и при повторном анализе страницы будет обрабатывать только те сообщения, которые не обрабатывались при

предыдущем обходе. Если на странице не появились новые сообщения, программа вернет пустой файл.

Контроль за ходом выполнения программы

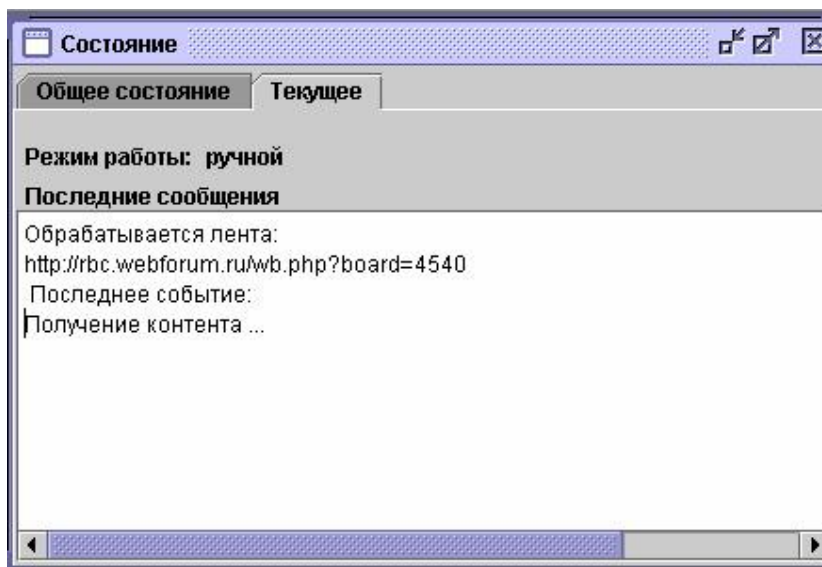
Для контроля за ходом выполнения программы служит окно "Состояние", в котором имеются две вкладки "Общее состояние" и "Текущее состояние".

"Общее состояние"



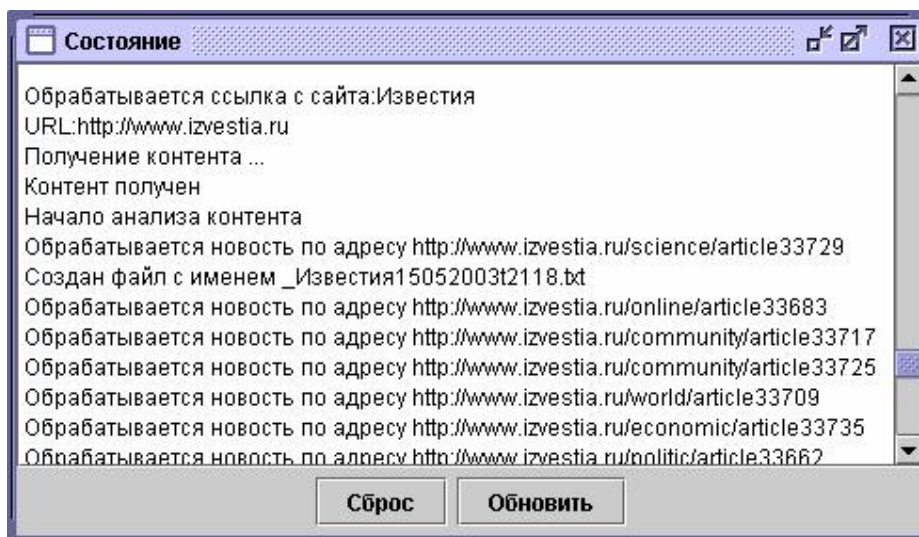
Указывает режим работы и отображает состояние каждой ссылки каждого сайта. Выбирая из списка сайт и его ссылку можно посмотреть, когда последний раз обрабатывалась ссылка, когда она будет обработана при условии работы по таймеру, сколько сообщений собрано при последнем обращении и за какое время. Отображаются сообщения об ошибках, имевших место при последней обработке.

"Текущее состояние"



Указывает режим работы и отображает последние события, произошедшие в системе, а именно, какие страницы сейчас обрабатываются.

Кроме того, для контроля программой ведется лог, в который записываются все события. Просмотреть его можно, выбрав меню Вид->Показать лог-файл



Кнопка "Сброс" стирает лог-файл, "Обновить" - загружает в окно лог заново.

Просмотр результатов сбора

Доступ к программе просмотра новостей осуществляется через меню *Вид* -> *Просмотр результатов* программы-паука, или запуском программы *Avalanche* через меню *Пуск (Start)*->*Программы (Programs)*->*Avalanche*->*Avalanche*.

Программа предоставляет большие возможности по поиску, фильтрации, организации, редактированию и экспорту сообщений, а также множество других функций.

Подробнее о программе *Avalanche* см. в разделе справки *Работа с поисковиком Avalanche*.

Горячие клавиши

Программа *Spider* допускает использование следующего набора стандартных горячих клавиш:

Ctrl + A Выделяет весь текст в окне

Ctrl + C Копирует выбранный текст в буфер обмена.

Ctrl + E Вставляет Internet-ссылку

Ctrl + V Вставляет содержимое из буфера обмена

Ctrl + X Вырезать выбранный текст в буфер обмена

Словарь терминов

Атрибуты сообщения	Характеристики и свойства сообщения, извлекаемого из сети Интернет, такие как дата, время, автор и источник сообщения.
Выделение сообщений	Распознавание поисковым роботом атрибутов сообщений на сайтах
Контент	Материалы, для размещения которых предназначен сайт (статьи, реклама, новости, каталоги, ПО, аудио и видео материалы и т.п.)
Лента новостей	Рубрика (страница) сайта, предназначенная для публикации новостей.
Новостная лента	См. «Лента новостей»
Паук	Поисковый робот – компьютерная программа, способная самостоятельно перемещаться по веб-пространству и извлекать необходимые пользователю материалы с просматриваемых сайтов.
База данных новостей	Каталог в файловой системе на ПЭВМ, в котором содержатся бинарные файлы с результатами работы программы
Ручное управление	Режим работы системы, когда запуск поиска осуществляется пользователем, при этом <u>паук</u> (см.) включается немедленно.
Скрипт дополнительной настройки	Текстовый исполняемый файл, содержащий инструкции по обработке сайтов.
Таймер	Режим работы системы, когда запуск поиска и формирование очереди для паука осуществляется программой с использованием таймера.
Тонкая настройка	Написание и ввод инструкций по обработке сайтов.
язык HTML	Hyper Text Markup Language - язык разметки гипертекста (текста содержащего ссылки, изображения и другие элементы), являющийся стандартом в среде World Wide Web сети Интернет. Имея документ в этом формате вы всегда сможете перенести его на другой компьютер и опубликовать в Интернет или интранет.